Final Report

Automated Performance Assessment and After-Action Review

Submitted to the Office of Naval Research

Insert Date

Contract Awards: N0001410IP20008; N0001411IP20012; N0001411IP20013;

Prepared By
Sandia National Laboratories

Address Correspondence to:
James (Chris) Forsythe
Sandia National Laboratories
MS 1188
Albuquerque, NM 87185-1188
jcforsy@sandia.gov

Executive Summary

The objective of the Automated Performance Assessment for After-Action Review project was to mature technical capabilities developed by Sandia National Laboratories for automated performance assessment to develop a product that could be transitioned to Naval aviation applications. Prior to this project, Sandia had demonstrated the feasibility of the AEMASE (Automated Expert Modeling and Student Evaluation) approach through a limited aviation-based demonstration. The AEMASE approach involves a three-step process. In the first step, individuals of desired levels of expertise demonstrate behavior on a simulator or within an instrumented environment. Second, the data generated through these demonstrations provide the input to machine learning algorithms that are used to derive a model of expert performance. Then, during training, real-time data is fed into the model which predicts the range of behaviors that would be expected within similar situations. The behavior of students may be compared to these predictions and discrepancies provide the basis for indentifying deficiencies in the knowledge or skills of individual students. These discrepancies are flagged on an after-action instructor debrief and allow instructors to focus after-action review on the specific needs of each student.

The AEMASE approach offers multiple benefits to Naval aviation. The use of machine learning approaches for deriving expert models streamlines the development process reducing costs by largely eliminating the costly and time-consuming steps of knowledge engineering required with most expert and intelligent tutoring systems. Second, during training, automated performance assessment provides continuous monitoring of students lessening the burden on instructors and allowing instructors to make better use of their time and resources.

Over the course of this four-year project, AEMASE was implemented with the E-2 Enhanced Deployable Readiness Trainer (E2EDRT) and automated measures developed to assess the performance of E-2 Hawkeye Naval Flight Officers. Laboratory experiments were conducted to first establish the accuracy of AEMASE-derived automated measures, as compared to manual graders, and second to demonstrate the utility of AEMASE-based automated performance assessment through improved training effectiveness. Subsequent work expanded the initial capabilities which were restricted to individual performance and employed only behavioral data from machine transactions to address verbal communications as one key facet of assessing team performance. Finally, using data obtained during large-scale joint forces exercises, the generalizability and scalability of AEMASE to fleet force-on-force exercises was demonstrated. At the conclusion of the project, the AEMASE capabilities are on track for transition to the Navy's E-2/C program as an upgrade to the E2EDRT.

Background

Prior to this project, Sandia National Laboratories had shown the feasibility of automated performance assessment tools such as the Sandia-developed Automated Expert Modeling and Student Evaluation (AEMASE) software. One technique employed by AEMASE is the grading of student performance by comparing their actions to a model of expert behavior. Models of expert behavior are derived by collecting sample data from simulator exercises or other means and then employing machine learning techniques to capture patterns of expert performance. During training, the student behavior is compared to the expert model to identify and target training to individual deficiencies. Another technique utilized by AEMASE is the grading of student performance by comparing their actions to models of good and/or poor student performance. Students with good and bad performance are identified and machine learning techniques are employed to construct models of these two types of performance in the same manner as expert performance. Student performance from other training sessions is then compared to these models to identify and target training to individual deficiencies. Both techniques avoid the costly and time-intensive process of manual knowledge elicitation and expert system implementation (Abbott, 2006).

In a pilot study, AEMASE achieved a high degree of agreement with a human grader (89%) in assessing tactical air engagement scenarios (Abbott, 2006). However, the 68 trials assessed utilized only four subjects with only three different training scenarios and the range of correct behaviors was quite limited.

Most research into automated student evaluation has been conducted in the context of intelligent tutoring systems. Murray (1999) provides a survey of intelligent tutoring systems, while Corbett (2001) provides a review of the empirical support for their effectiveness. Jensen, Chen, and Nolan's (2005) work on Combined Arms Command and Control Trainer Upgrade System (CACCTUS) provides one exception. This tool analyzes events from training sessions to find causal relationships among student errors and undesirable outcomes. The system then applies a set of rules to determine and highlight the correct behaviors. This work differs from AEMASE in that AEMASE attempts to learn a model for correct behaviors by observing experts, instead of relying on a crafted rule base. Relatively few efforts have sought to automatically acquire models of correct behaviors. Anderson, Draper and Peterson (2000) used neural networks to create behavioral clones for piloting simulated aircraft, but their work focused on personal insights based on examination of neural network models of individual students. AEMASE uses its learned models to compare novice and expert behavior automatically.

Problem

The U.S. armed services are widely adopting simulation-based training, largely to reduce costs associated with live training. Ideally, instructors would observe each individual within the context of team performance and provide instruction based on observed misunderstandings, inefficient task execution, ineffective or inappropriate actions and so forth. However, it is impossible for instructors to devote this level of attention and time to each student. To maximize training efficiency, new technologies are required that assist instructors in providing individually-relevant instruction.

A significant cost in simulation-based training is the workload on human instructors to monitor student actions and provide corrective feedback. For example, the U.S. Navy trains Naval Flight Officers for the E2-Hawkeye aircraft using a high-fidelity Weapons Systems Trainer (E2 WST). Currently this requires a separate instructor to observe each student within the context of team performance. Individualized instruction contributes to high training costs. Intelligent tutoring systems target this need, but they are often associated with high costs for knowledge engineering and implementation. New technologies are required that assist instructors in providing individually-relevant instruction.

Objective

The objective for this project was to extend technologies for Automated Expert Modeling and Student Evaluation (AEMASE) to a platform that is relevant to Naval aviation training, and conduct research to establish the accuracy of automated student assessments and the training benefit achievable with this technology. In particular, AEMASE was integrated with the E-2 Hawkeye Enhanced Deployable Readiness Trainer (E2EDRT), and development and testing has focused on advancing capabilities for training E-2 Naval Flight Officers.

Approach

The goal of AEMASE is first to let subject matter experts rapidly create and update their own models of normative behavior and then use these models to evaluate student performance automatically (Abbott, 2006). The system operates in three steps. First, the system must acquire examples of behavior in the simulated environment. Next, machine-learning techniques are used to build a model of the demonstrated tactics. The system then compares student behaviors in the same task environment to the expert model to establish a score. Afterwards, the student and instructor can review the training session by interacting with a plot of the time-dependent grade. The remainder of this section provides additional detail on these steps.

In the initial step, the system records examples of task behavior. The examples may include both good and bad behavior performed by either students or subjects matter experts. Examples may be obtained by performing exercises on the target simulator or within a relevant proxy environment. However, a subject matter expert must accurately grade the examples to provide AEMASE with points of reference in its comparisons to student behaviors during evaluation. After acquiring graded example behaviors, the system applies machine learning algorithms to create the behavior model. An appropriate learning algorithm must be selected for each performance metric, depending on the type and amount of example data available, such that the resulting model generalizes assessments of the observed behaviors to novel student behaviors. We have implemented a suite of machine learning algorithms (e.g. neural networks, instance based/ nearest neighbor algorithms, support vector machines, linear regression, rule induction) and cross-validation tests to determine which algorithm makes the most accurate predictions for each metric. Finally, the system uses the learned behavior model to assess student behaviors. As each student executes a simulated training scenario, his or her behavior is compared to the model for each performance metric. The model determines whether student behavior

is more similar to good or bad behavior from its knowledge base, and helps to identify and target training to individual deficiencies. Initially, the knowledge base is sparse, and incorrect assessments may be common. However, the instructor may override incorrect assessments. The model learns from this interaction and improves over time. For the research described here, we used AEMASE as a tool for after action reviews (See Figure 1), although the system could also be used to provide students with feedback throughout a training exercise.

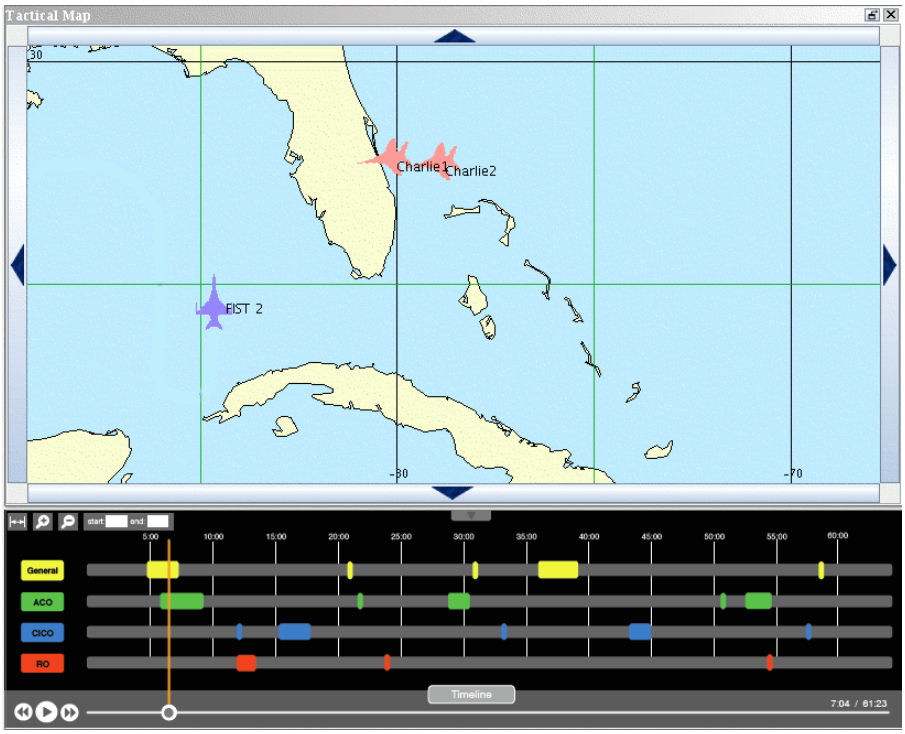

Figure 1. Debrief Tool With Automated Event Flagging. The debrief tool displays a video replay of the operator console (similar to this map display), and a timeline of events suggested by AEMASE for discussion during debrief. The tool also includes visualizations of entity movement over time.

For training Naval Flight Officers, we used two basic types of AEMASE metrics. The first type of AEMASE metric is Context Recognition, which assesses whether the student is maintaining the tactical situation within norms established by previous expert demonstrations. This is done by monitoring the values of one or more continuous metrics (e.g. positions, ranges, headings, fuel load, etc). Unexpected combinations of values indicate the student may not know what to do, or may be losing control of the situation. The Fleet Protection metric described below is a simple (one-metric) example.

The second type of AEMASE metric is Sequence Recognition, which assesses whether certain sequences of events provoke the expected sequence of responses. An example is Labeling Neutral Entities; a set of events (appearance of a radar track, detection of certain RF emissions) should lead to specific actions by the subject (labeling the track as a non-combatant). Any failure of the student to complete the sequence within a time limit (determined by modeling expert response times) is flagged for review.

AEMASE Automated Performance Assessment Accuracy

Establishing the validity of automated assessments requires studies in a realistic training environment, rather than just a simple laboratory task. E2 operators are trained and tested on several different simulators ranging from a part-task computer-based training (CBT) system that runs on a single PC, to the high-end E2 WST system which faithfully replicates most aspects of E2 operations (ranging from the physical controls to system fault diagnosis and recovery) and requires a team of instructors and operators to conduct training. For this study we used the E-2 Distributed Readiness Trainer (E2EDRT), a medium-fidelity trainer which presents students with the same mission software used on the E2 aircraft. Multiple instructors are needed to evaluate simulation training and sessions can last hours at a time. Automated assessment of E2 operator performance in these sessions would greatly reduce instructor workload and would increase overall efficiency.

Participants

Twelve employees from Sandia National Laboratories volunteered to participate in the experiment. The participants met certain required criteria for the experiment which reflected the requirements for an entry-level E2 Hawkeye operator. In addition, two former E2 Hawkeye operators served as subject matter experts (SME's).

Materials

Materials included an E2 Deployable Readiness Trainer (EDRT) simulator that was obtained from the Naval Air Systems Command's Manned Flight Simulator organization. The Joint Semi-Automated Forces (JSAF) simulation software was used to create and drive the training and testing scenarios.

Procedures

The participants were recruited via an advertisement and those who responded positively and met the required criteria were included in the study. The participants were scheduled for an initial all-day training session in which a former E2 Hawkeye Naval Flight Officer provided a tutorial on E2 operations emphasizing the basic radar systems task that would be the subject of the experiment. The participants were also asked to sign an informed consent. After the initial training session, the participants were scheduled for seven additional training sessions. The participants were lead through the sessions in the same order. Once they had finished the training sessions, the participants completed two testing sessions. The participants completed the seven training and two testing sessions individually.

Training Sessions

The first five sessions consisted of additional training sessions designed to teach the participants the basic operations of the E2 radar system in depth on the E2EDRT. For each session, the experimenters first demonstrated the proximate operation(s) on the E2EDRT and then the participant was asked to perform the operation(s) in scaled down, yet realistic, simulations. Since all five of these sessions were for training purposes, the experimenters were available to answer questions. At the end of each training session, the participants filled out a questionnaire indicating their understanding of the operation(s) on the preceding training session. At the end of the fifth scenario, the participants

completed a questionnaire assessing their knowledge of all of the operations learned in the training sessions.

Testing Sessions

The last two sessions were testing sessions in which the participants were assessed on their knowledge of the operations and tactics covered in the five training sessions. The participants completed these more difficult simulations without the help of the experimenters. At the end of each testing session, the participants were asked to complete a questionnaire which queried their confidence in their performance on the preceding test scenario.

Metrics

Based on guidance from the SMEs, three metrics were developed which were used to grade the participants' performance in the testing sessions.

Fleet Protection - Participants were instructed to prevent non-friendly entities from nearing the carrier group. The amount of time the non-friendly entities were within a pre-specified proximity of the carrier group was assessed.

Labeling Neutral Entities - Participants were instructed to promptly and appropriately label any neutral entity that appeared on the radar scope. The latency with which the participants took to label these entities was assessed.

Battlespace Management - Participants were instructed to effectively manage their air assets as the battle space evolved during the scenario. This included re-positioning CAP (Combat Air Patrol) stations so that friendly airspace would not be violated.

Assessments

Manual Assessments

Two trained experimenters independently reviewed video recordings of each of the testing scenarios for all participants. The experimenters graded the participants' performance on the three metrics for the two testing scenarios. For each metric, the two experimenters specified instances of good and instances poor student performance. These instances formed subsets of manual assessment data that were used in training the AEMASE automated performance measures

Automated Assessments

The participant performance on the two testing scenarios was assessed by AEMASE. AEMASE used the good and poor instances indentified by the two experimenters as base examples from which to assess participant performance.

Results

The manual assessments and the automated assessments were compared for each of the three metrics.

Fleet Protection

Manual assessment was based on the amount of time the non-friendly fighters spent within too close of a proximity to the carrier group. The inter-rater reliability between the two experimenters was 99%. The automated assessment used a proxy measure, which consisted of the distance between the carrier group and the closest non-friendly asset. The results indicate a 100% agreement between the automated and manual assessments in terms of identification of unsatisfactory student performance (i.e., periods in which students allowed non-friendly assets to get too close to the carrier group).

Labeling Neutral Entities

Manual assessment was based on reviewing the time-stamped recording of when the neutral entities were labeled. The inter-rater reliability between the two experimenters was 94%. The automated assessment was based on the analysis of network messages from the mission computer. The results indicate a 95% agreement between the automated and manual assessment for correct labeling of the neutral entities.

Battlespace Management

Manual assessment was based on the time and accuracy with which the CAP stations were re-positioned. The inter-rater reliability between the two experimenters was 99%. The automated assessment was based on post-hoc analysis of radio communications. Results indicate an 83% agreement between the automated and manual assessment.

AEMASE After-Action Debrief Utility

A significant cost in simulation-based training is the time demands on human instructors who monitor student actions and provide corrective feedback. The work presented here focuses on U.S. Navy training of Naval Flight Officers for the E-2-Hawkeye aircraft using a high-fidelity simulator. The three flight officers must learn to detect, track, and identify all assets, such as aircraft, and to provide communication among the commanding officers and all friendly assets. This currently requires a separate instructor to observe each student within the context of team performance and provide instruction based on observed misunderstandings, inefficient task execution, and ineffective or inappropriate actions. Such individualized instruction is labor intensive and contributes to high training costs. The purpose of this study was therefore to determine whether a group given verbal feedback from an instructor on their performance using an AEMASE-based debrief tool would outperform a group given verbal feedback alone.

Participants

Volunteer civilian employees were recruited via advertisement. All twenty-two participants met criteria for the experiment that reflected the requirements for an entry-level E-2 Hawkeye operator. The participants were both men and women and were between the ages of 20 and 28. The participants were split into two groups: a control group (N=12) and a debrief group (N=10). Two experienced E-2 Hawkeye Naval Flight Officers served as subject matter experts.

Materials

The materials and equipment used in this study were the same as that employed for the previous study. In particular, this included the E2EDRT and JSAF simulation exercises.

Procedure

The participants provided informed consent and were then scheduled for an initial eight-hour training session. Here, an E-2 Hawkeye Naval Flight Officer provided a tutorial on E-2 operations emphasizing the basic radar systems task that would be the subject of the experiment. Following this initial session, the participants were scheduled individually for five simulation-based training sessions. All participants were led through these sessions in the same order. After finishing the training sessions, the participants individually completed two testing sessions. Two trained experimenters graded each participant's performance and performance was compared between the two groups.

Training Sessions

The five simulation-based training sessions were designed by an E-2 subject matter expert to teach the basic operations of the E-2 radar system on the simulator. The topics included simulator familiarization, check-in procedures, and managing air assets, managing surface assets and integration of air and surface pictures in complex tactical scenarios. For each session, the experimenters first demonstrated the proximate operation(s) on the simulator, after which the participant was asked to perform the operation(s) in scaled down, yet realistic, simulations. Since all five of these sessions were for training purposes, the experimenters were available to answer questions. Each training session lasted approximately 1.5 hours.

For the control group, the instructor gave participants real-time, verbal feedback of their training session performance deficiencies. For the debrief group, the instructor used a debrief tool featuring graphical depictions (e.g., timeline and occupancy maps derived by AEMASE) of participants' performance in addition to real-time, verbal feedback. The instructor was given sufficient training on how to use the debrief tool before the experiment started.

Testing Session

The last two sessions were testing sessions in which the participants were assessed on their knowledge of the operations and tactics covered in the five training sessions. The participants completed these more difficult simulations without the help of the experimenters. Each testing scenario lasted about 1 hour.

Metrics

The selected metrics correspond to a subset of those used by the Navy in training Naval Flight Officers, and included fleet protection, labeling of neutral entities, and battlespace management.

Fleet Protection

Participants were instructed to prevent non-friendly entities from nearing the carrier group. Performance was assessed based on the latency to commit friendly fighters to enemy fighters as they approached the carrier group. During training, participants were given feedback regarding how quickly

they committed friendly fighters to non-friendly entities entering the battlespace. For those in the debrief condition, the Debrief tool was used to playback the scenario (during training) and participants were shown their performance.

Labeling Neutral Entities

Participants were instructed to label any neutral entity that appeared on the radar scope promptly and appropriately. This required a high degree of situational awareness due to the large number of radar tracks. The complexity of a scenario also prompted subjects to fixate on a small portion of the battlespace. The accuracy and latency with which the participants labeled these entities was assessed. During training, participants were given feedback regarding how quickly and accurately they labeled neutral entities. For those in the debrief condition, the Debrief tool was used to playback the scenario in order to point out the participants' mistakes.

Battlespace Management

In one test scenario, the student was instructed to re-task fighter aircraft away from the initial combat air patrol station. Moving the fighters created a gap in air defenses, possibly allowing an incursion into protected air space as shown in Figure 2. The student was expected to notice this vulnerability and re-assign other fighter assets to fill the gap.

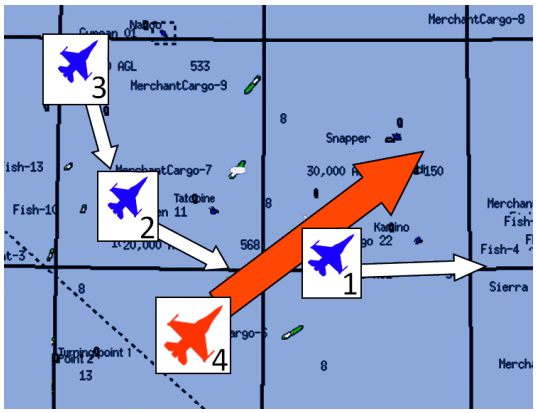

Figure 2: Battlespace Management. In this battle problem, Fighter 1 is re-assigned to the East, leaving a gap in air defenses. The student should move Fighters 2 and 3 to fill the gap; otherwise, enemy Fighter 4 may penetrate the defenses.

At this time, AEMASE could not recognize speech from radio calls, so the automated assessment was based on analysis of readily available simulation data, such as the positions of friendly and enemy fighters over the course of the scenario. One method used to represent this data was an Occupancy Grid, shown in Figure 3. The battlespace was divided into a grid and the total amount of time spent in each grid cell by friendly and enemy fighters was computed, resulting in two matrices of time-weighted values. This approach is more informative than simple "snail trails" left behind by each entity because it captures information about how much time an entity spends at a location.

During training, participants were given feedback regarding whether or not they correctly re-tasked friendly fighters. Those in the debrief group were also shown how their AEMASE Occupancy Grid differed from an expert's Occupancy Grid (Figure 3).

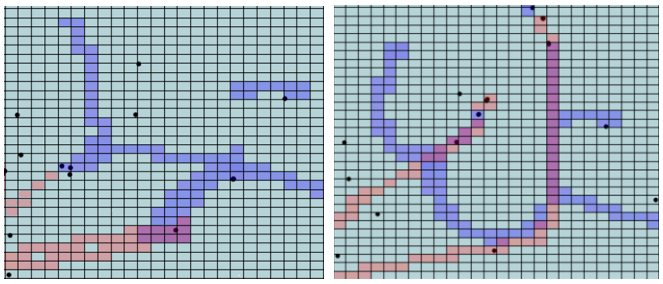

Figure 3: Occupancy Grids. Blue and red tracks show the paths of friendly and opposing forces, respectively. On the left, friendly forces were pre-positioned correctly and repelled the incursion. On the right, gaps in defenses allowed the penetration of protected airspace.

Results

The first metric concerned fleet protection. In this case, students were evaluated with regard to their allowing enemy aircraft to approach the carrier group, and specifically, the timeliness with which subjects committed friendly aircraft to intercept hostile aircraft posing a threat to the carrier group. As illustrated in Figure 4, subjects in the debrief condition performed significantly better than those in the control group ($t=2.03$, $p<0.05$). On average, these subjects executed radio calls committing aircraft prior to hostile aircraft penetrating the commit line, whereas the control group, on average, did not execute radio calls until after hostile aircraft had passed the commit line.

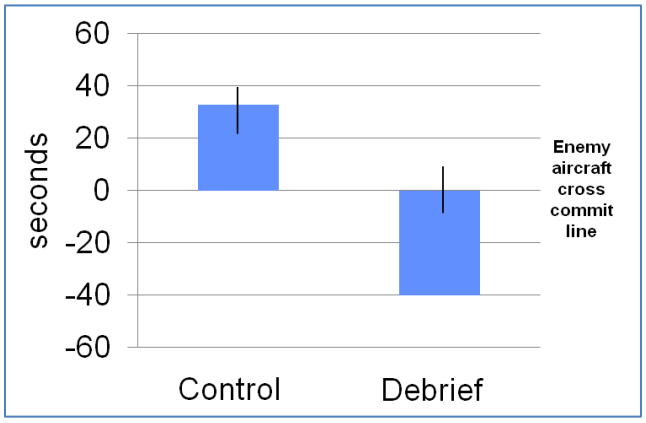

Figure 4. Subjects trained using the AEMASE debrief tool executed radio calls committing friendly air assets to intercept hostile aircraft posing a threat to the carrier group significantly earlier than subjects in the control group.

The second metric involved the timing and accuracy of labeling commercial aircraft. Subjects trained using the AEMASE after-action review labeled aircraft significantly sooner ($t=1.69$, $p<0.05$) and more

accurately (t=1.87, p<0.05) than subjects in the control group (See Figure 5). It is emphasized that performance was superior for both speed and accuracy, implying that subjects had an overall better appreciation of the requirements of the task and situation awareness.

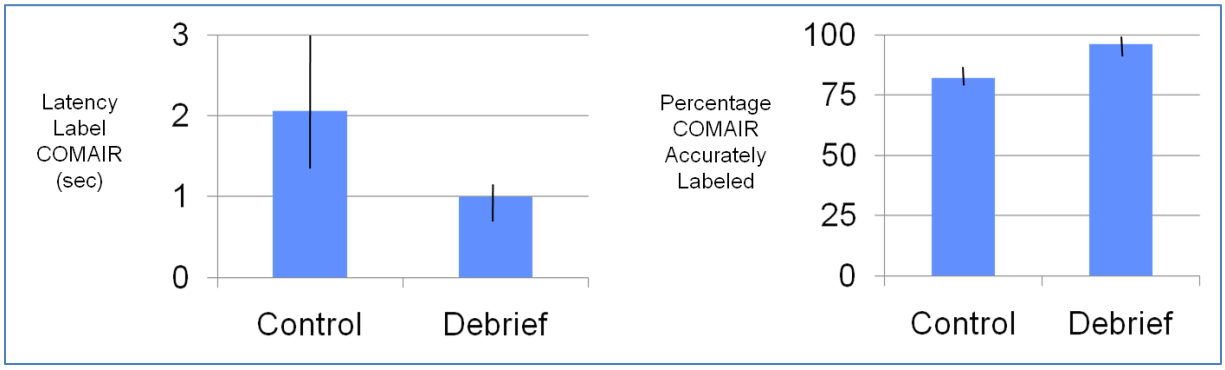

Figure 5. Subjects trained using the AEMASE debrief tool labeled commercial aircraft sooner and more accurately than subjects in the control group.

The third metric, focused on battlespace management, and a specific aspect of asset management whereby students must recognize a gap in their air defenses and re-position Combat Air Patrols accordingly. There was no difference between the experimental and control group for this metric, and in actuality, only three of the twenty-two subjects performed this tactic correctly. In retrospect, it was concluded that this tactic was conceptually too advanced given the limited training subjects received.

In addition to the three planned comparisons, a fourth metric considered the timeliness with which subjects informed the warfare commander after receiving reports that enemy aircraft had been shot down. Subjects trained with the AEMASE after-action review reported kills significantly sooner (t=2.66, p< 0.005) than subjects in the control group (See Figure 6).

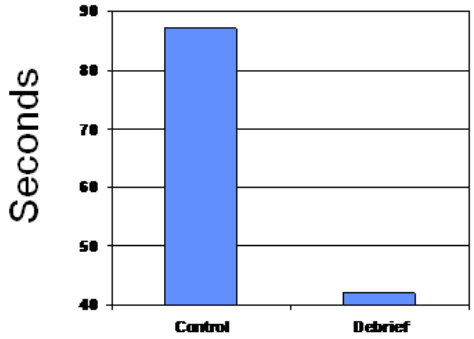

Figure 6. Subjects trained with the AEMASE debrief tool informed the warfare commander significantly sooner after receiving reports that enemy aircraft had been shot down.

Automated Assessment of Team Performance

Initial implementations of AEMASE focused on training individual skills. A primary research objective has been to extend these capabilities to enable automated performance assessments for teams. Per guidance from Navy representatives, the Team Dimension Training (TDT) paradigm was adopted as a conceptual model of teamwork from which metrics could be derived. A challenge with automated performance assessment is that data for metrics must be available within the system, or attainable through additional instrumentation. This consideration has significantly constrained the potential metrics achievable with the E-2 Enhanced Deployable Readiness Trainer, given the requisite data for most measures of teamwork is unavailable. Consequently, attention in developing team metrics has been focused on speech communications and labeling of entities.

E2 subject matter experts identified radio communications as a critical aspect of E-2 training. This poses a challenge for automated performance assessment because current technologies for automated speech recognition and natural language processing have limited capabilities. Rather than attempting to solve the well-studied problem of accurate speech recognition, an approach was employed that combined approximate speech recognition with tactical context recognition. For example, if an allied fighter is intercepting an enemy fighter, the E-2 NFO should use appropriate TACAIC terminology to enhance the situational awareness of the allied aircraft. Without assuming perfect speech-to-text conversion, it may still be possible to determine when an air engagement is taking place, and whether the NFO is using TACAIC terminology. The first step in this research was to identify the best commercially available voice technologies.

There are many factors which contribute to the contextual appropriateness of speech communications, including both non-linguistic (e.g., duration and frequency of utterances) and linguistic factors. Since perhaps the most important linguistic factor is the category of utterance (e.g. ABCC, TACAIC) (that is, was the category of utterance appropriate at the time?), our basic approach to evaluation was to consider voice recognition as part of a larger system in which the ultimate metric is the categorization accuracy. This contrasts with an approach which considers the number or percentage of words correctly recognized; the latter approach fails to take into account that consistent misrecognition of a word or phrase may have no negative impact on the categorization of the utterance. The basic architecture envisaged is depicted in Figure 7 below.

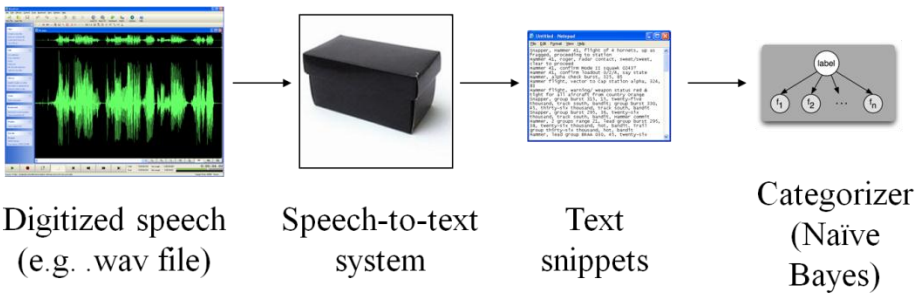

Figure 7. Proposed architecture employing commercial speech recognition as a basis for categorizing speech utterances for automated performance assessment.

For speech-to-text, a survey of 18 open-source and commercial off-the-shelf voice technology systems was conducted. A number were quickly eliminated for reasons including domain specificity (e.g. exclusively for medical industry) and vendor viability (e.g. product line had been discontinued). After down-selecting, we were left with two packages: Microsoft's Speech API 5.1 and Nuance's Dragon NaturallySpeaking. We found no statistically significant difference between the categorization accuracy for the two systems: 56% for Microsoft versus 58% for Nuance. Note, these percentages reflect accuracy in placing utterances in one of seven categories (See Table 1). For this assessment, software was deployed in speaker-independent mode (i.e. no individualized training), the grammar used was based on word probabilities, and categorization was based on whole utterances and recognition as opposed to hypothesized speech. Other approaches were examined that included alternative grammars and different settings within grammars regarding the grammar rules and phrase probabilities, but these approaches were less effective with categorization performance ranging between 0-38%. Figure 8 summarizes the results for each category of utterances showing that accuracy varied substantially across the categories. The baseline accuracy is about 15%, so it is clear that incorporating either SAPI 5.1 or Dragon into the larger system could be a viable approach.

Table 1. Categories of speech utterances identified for E-2 Hawkeye NFO operations and used in assessing categorization accuracy of commercial speech recognition.

- Administrative check-in of assets, e.g. "Hammer flight, vector to Cap station alpha, 324, 83"
- TACAIC, "Snapper, group burst 295, 36, twenty-six thousand, track south, bandit, Hammer commit"
- Communications with AW, e.g. "AW, T, Hammer flight committed track number 4701"
- Communications with AZ, e.g. "AZ, T, track number 4715 identified as hostile surface action group"
- SAR, e.g. "Sierra, be advised, your pony requested to conduct search and rescue operation for downed aircrew of Hammer 44. Say status your pony"
- ABC2, e.g. . "Hammer 41, copy airborne from Bagram, take angels 35, vector to stack 045, 95"
- recognized, as opposed to hypothesized speech

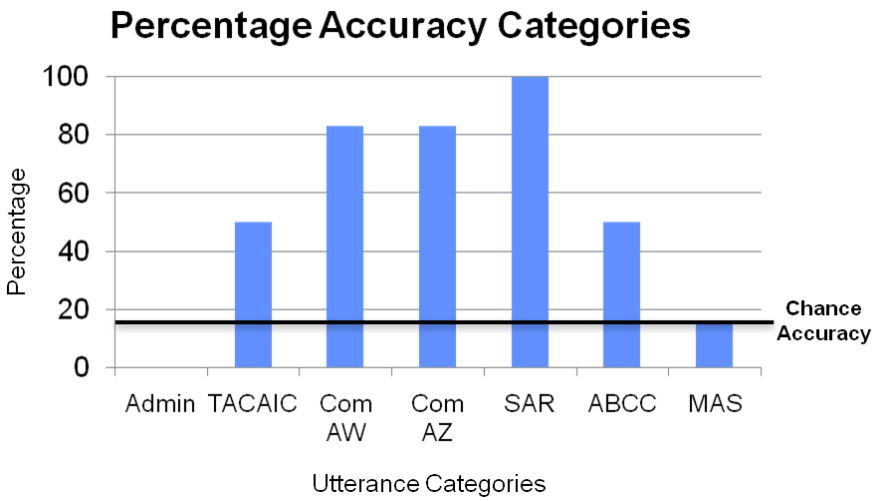

Figure 8. Combined accuracy for each category of utterance, with the likelihood of correct categorization by chance being 15%.

The next step in employing speech-based communications as a basis for automated performance assessment involved analysis comparing the speech patterns of expert and novice teams. In this study, two two-person teams consisting of expert E-2 NFOs and two two-person teams consisting of novices (i.e. test subjects from the previously discussed project that had attained a modest level of individual proficiency completing scenarios on the E2EDRT) completed scenarios on the E2EDRT. The speech of both expert and novice teams were recorded. We hypothesized that the language of the teams would be useful in discriminating between experts and novices. This approach was inspired by earlier research in which TF/IDF (term-frequency/inverse-document-frequency) with Latent Semantic Analysis was highly effective in automated essay grading, despite disregarding the order of word usage (Foltz, Laham & Landauer, 1999). Our primary concern concerned whether tf/idf would be effective given the limited accuracy of automated speech recognition.

In the development of speech-based team performance measures, it was necessary to work with off-the-shelf commercial speech recognition software, which our benchmark testing indicated would only provide an approximately 60% level of accuracy for speech-to-text transcription. Consequently, it was necessary to select metrics that were suitable given the available capabilities of the speech recognition software. Based on interaction with subject matter experts (reservist E-2 NFOs), three aspects of team communication were identified: (1) when a team communicates, (2) what they communicate, and (3) how they communicate. By studying when NFO's communicate, the responsiveness of the team to external events and information flow within the team may be assessed. However, what NFO's communicate is just as important — each utterance should transmit important information communicated in a clear and understandable manner. Finally, the phonetic characteristics of the communication (e.g., tone and rhythm) play an important role in conveying cues such as urgency or importance.

It was found that experts and novices differ significantly in how often they communicate, with expert teams making many fewer radio calls of less duration than both novice teams (Figure 9). This finding is consistent with the notion that given limited radio bandwidth, experts learn to conserve bandwidth by communicating only when it is necessary and communicating in a concise manner.

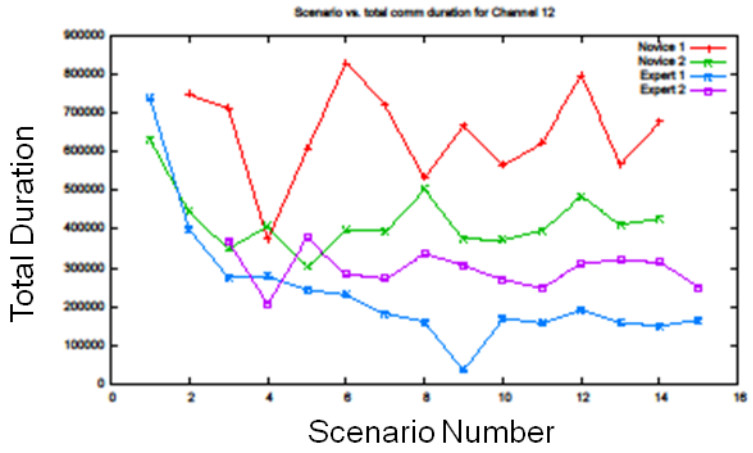

Figure 9. The total duration of radio calls for expert (blue and purple lines) and novice (red and green lines) for E-2 NFO teams across the sixteen test sessions.

Next, the semantic content of utterances was considered. For two NFO's to effectively communicate they must share a common language. Using speech-to-text conversion, the resulting text documents were compared using term frequency-inverse document frequency (tf-idf) methods. As shown in Figure 10, for 7 of 8 subjects, the two most similar subjects (using cosine-similarity of term vectors) were in the same category of expertise (novice/expert). This means that the semantic content of utterances by experts was more similar to other experts, than novices, and vice versa.

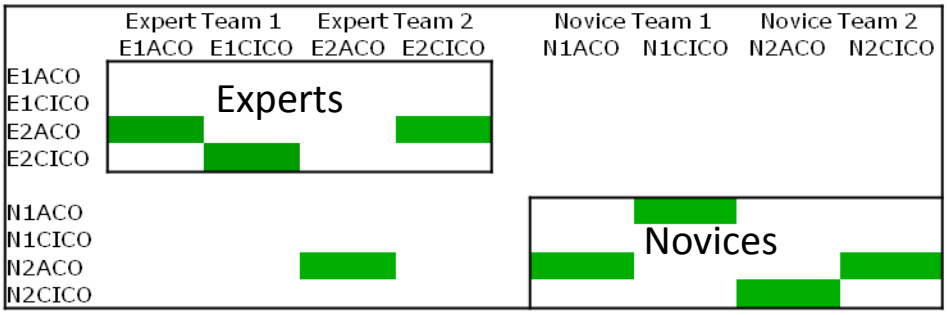

Figure 10. Semantic analysis of radio communications showed experts were more similar to other experts, and novices to other novices (green indicates the subject that each subject had the greatest similarity).

More detailed analysis revealed that there was a distinct difference in the word usage for experts as compared to novices. In general, there was a specific vocabulary used by both expert and novice teams, yet the experts showed little variance from this vocabulary, as compared to the novices. One observation concerned the use of filler words (e.g. “um,” “ er,” etc.) As shown in Table 2, the prevalence of filler words was much greater for novice than for expert teams. This finding is not surprising given that filler words serve a purpose in the turn-taking that occurs with verbal discourse. By using a filler word when an individual needs time to formulate the words they want to speak, they sustain their turn, buying time to search memory and construct their utterance.

Finally, initial analysis indicated that experts may be distinguished on the basis of phonetic properties of their verbal communications (e.g. situationally-appropriate urgency). A detailed analysis was undertaken using frequency components of speech with particular attention focused on periods in the experimental test scenario immediately preceding and following commencement of hostilities (i.e. the scenario went “hot”). These analyses failed to identify characteristics that would allow the experts to be distinguished from the novices.

Table 2. The frequency of filler words during an illustrative scenario demonstrates the greater reliance on filler words by novice teams.

	Experts	**Novices**
ah	1	6
er	4	8
like	5	9
uh	112	307
um	5	28

Based on this research, it was concluded that speech-based measures provide a viable method for distinguishing expert from novice teams of E-2 NFOs. While the quality of speech communications represent only one facet of an individual's capacity to work effectively within the context of a team, as emphasized within the Navy's Team Dimension Training (TDT), speech quality is vital for effective team performance.

Applied as a basis for targeted training through AEMASE, two measures emerge that would be informative during after-action debrief. The first would involve flagging periods in which the frequency and duration of student utterances deviate from the predictions of an expert model. This would allow instructors to identify situations in which students either communicate too little or too much. One example would involve air-to-air engagements during which fighter pilots expect a continuous update from the E-2 NFO concerning the hostile forces. An AEMASE implementation may be envisioned in which instructors are provided flags on the AEMASE timeline indicating points at which the student NFO did not provide pilots the desired frequency of verbal updates.

Second, an implementation may be envisioned that flags periods in which student NFOs utter an excessive frequency of filler words. Given a key facet of E-2 NFO training involves teaching students the appropriate vernacular, this measure would indicate to instructors points at which students understand that they need to say something, but are unsure of the correct words.

AEMASE After-Action Debrief

The product of this project to be delivered for transition to the Naval aviation enterprise is the AEMASE after-action debrief. This software product consists of an after-action review capability featuring AEMASE automated performance assessments. In operation, as students complete an exercise, the exercise will be recorded. During this time, instructors may insert flags on a timeline to signify events for consideration during after-action review. Once the exercise has completed, the timeline will be populated by both the flags that instructors have manually entered and flags automatically inserted by AEMASE. The instructor may then use the timeline to navigate to different points within the recorded exercise and playback periods of interest.

The initial implementation of the AEMASE after-action debrief will occur as an upgrade to the E2EDRT. This will occur through integration of the AEMASE after-action debrief with the Navy's Common Distributed Mission Training Systems (CDMTS). This integration was accomplished and development undertaken to provide the features deemed essential by fleet representatives for an effective E-2 after-action debrief. Key facets of this development included the following.

Data Recording

Recording consists of video capture, voice capture, HLA network data capture, operational flight program capture, and system integration of these components.

Video Capture records the students' ACIS (Advanced Control Indicator Set) screens. Video capture was implemented using the RGB DSx 200 CODEC (compressor/decompressor) devices which create an H.264 compressed video stream and send it over the local network to the recording computer.

Voice Capture records the students' radio calls. These are used during replay (for students to hear their own performance) and for voice-based metrics as documented above. For voice capture, network data traffic from the ASTi radio simulator used in the EDRT is recorded. Capturing this data required protocol conversion from DIS (Distributed Interactive Simulation) to HLA (High-Level Architecture). This is done using the JLVCDT (Joint Live/Virtual/Constructive Data Translator) software which we configured with assistance from Alion Science and Technology Advanced Modeling & Simulation Technologies Operation. The Sandia-developed HLA Data Logger was then upgraded to receive information from JLVCDT without re-sending it to the NCTE HLA federation by creating a private HLA federation that exists only within the AEMASE Data Logging computer itself.

HLA Network Data Capture was enhanced to avoid requesting any unnecessary information from the NCTE network, which could increase network traffic. Sandia software was modified so it only requests the same information as the EDRT. This was non-trivial because the EDRT only requests information within a specified radius of the simulated E-2C entity, which moves over time. Thus our logging software must communicate with the EDRT during the exercise to update its DDM subscriptions. This was accomplished through a shared memory mechanism used by the EDRT, which was implemented in the Sandia logging software.

Operational Flight Program Capture records the student actions on the training system, such as labeling radar tracks as friendly, neutral, hostile, etc. This is done by monitoring the network messages between components of the computer system aboard the E-2C (and in the EDRT). This was implemented by configuring the CISCO router in the EDRT to include the AEMASE recording computer in the network where the mission computer sends information to the ACIS terminals.

System Integration consisted of configuring a new AEMASE Data Capture Host for the EDRT, creating a user-friendly GUI to control all of these recording components, and testing to verify correct operation in a simulated operational setting. Three rounds of tests were performed in conjunction with the Navy Manned Flight Simulator organization on their E2EDRT to verify the data capture functionality. These tests were successful and confirmed that the Sandia components did not degrade the operation of the E2EDRT or request extra information from the NCTE.

Replay

The main replay components are Video Replay, the Timeline display, and CDMTS for displaying HLA data and playing radio calls. These are integrated by software that synchronized playback among the components during debrief.

The MPlayer was adapted for video replay. This is a video player available for both Windows and Linux computers. It was chosen because it correctly replays videos from the RGB Spectrum DSx200 codec, and because it can be controlled for synchronized replay.

The Timeline display, which was already used in CDMTS, was updated to display scenario events from the AEMASE performance measurement database. This display shows manual annotations created by the instructor during the exercise as well as automated annotations created by AEMASE. Editing an annotation during debrief updates the annotation in both the Timeline display and in CDMTS (which has an embedded timeline display).

CDMTS is used by the AEMASE E2EDRT Debrief to replay HLA data so the user can inspect the recorded state of all simulation entities. A plug-in to CDMTS was developed which allowed use of the EDRT Debrief Synchronization Library. The other major CDMTS enhancement was a user interface for controlling replay of the recorded radio calls from the EDRT. This interface was non-trivial because the E2EDRT simulates 6 or more radios (depending on configuration).

Operational Test and Evaluation

The initial operational use of the AEMASE debrief system will be to support the EDRT in Fleet Synthetic Training (FST) events. These are large-scale networked events where the EDRT is connected to the NCTE (Navy Continuous Training Environment). Access to the NCTE is strictly limited, and there is no NCTE connectivity at Sandia Labs, so integration and test events were conducted offsite. We established a two-step procedure: first, initial testing at Manned Flight Simulator (MFS) at NAS Patuxent River, then final testing at NS Norfolk on an operational EDRT. Although MFS is not NCTE-capable, initial testing there allowed the government engineering team to verify the software, to reduce the technical risk of connecting new software to the NCTE.

The first round of testing was for the data recording software. (Development of the replay and analysis software could not be completed until after this sample data was collected for requirements development and testing). The Navy Warfare Development Command (NWDC) took the leading role in developing the test plan for this test. The test plan is included as Appendix I. The primary technical risk posed to the NCTE by the recording software was that the HLA logger might subscribe to all HLA messages instead of just the ones needed for the EDRT, which could overload the NCTE and disrupt an entire large-scale exercise. During the integration and test event we configured the HLA logger to capture only the necessary data, and NWDC personnel verified that running the HLA logger in this configuration placed no additional network load on the NCTE. The other main objective was to verify that all data required for replay was captured. The main issues were:

- Configuring the Logging software to capture all pertinent HLA and DIS network traffic
- Load testing the logger to verify it can handle the number of entities and update rate of the NCTE
- Endurance testing the logger to verify it can capture a 4 hour scenario

Several technical issues were found and corrected during the integration. After this, the test was successful.

The second (and final) round of testing will be to test the replay and analytic capabilities of the system, for which development has continued, enabled by the test data gathered in the first round. The second round of testing is planned for early 2012. The main challenge in conducting this testing is being granted access to the NCTE, which is a production environment in heavy use; thus opportunities for integration and test are relatively rare.

Conclusion

While the final steps leading to product transition are still underway, this project has successfully accomplished its stated objectives. In particular, research and development was undertaken to mature the Sandia AEMASE capability for automated performance assessment to the point that it could be incorporated within an operational simulation-based training system, specifically the E2EDRT. This research and development involved experimental studies that empirically established the accuracy of the AEMASE automated performance measures and demonstrated quantifiable improvements in training effectiveness. Additionally research quantified the performance achievable with off-the-shelf, readily available voice recognition software and demonstrated mechanisms by which this technology could be effectively employed as a basis for automated performance measures. Based on the results of this project, a capability has been developed that should serve the Navy by both reducing the costs of system development, while lessening manpower costs associated with Naval aviation training.

Acknowledgement

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Abbott, R. G. (2002). Automated tactics modeling: Techniques and applications. Dissertation Abstracts International, 68.

Abbott, R. G., (2006). Automated Expert Modeling for Automated Student Evaluation. Intelligent Tutoring Systems, 1-10.

Anderson, C. W., Draper, B. A. & Peterson, D. A. (2000). Behavioral cloning of student pilots with modular neural networks. In ICML ’00: Proceedings of the Seventeenth International Conference on Machine Learning, (pp. 25-32). San Francisco: Morgan Kaufmann.

Corbett, A. T. (2001). Cognitive computer tutors: Solving the two-sigma problem. In UM ’01: Proceedings of the 8th International Conference on User Modeling, (pp. 137 – 147). London: Springer-Verlag.

Foltz, P. W., Laham, D., Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. Proceedings of EdMedia ’99.

Jensen, R., Nolan, M. & Chen, D. Y. (2005). Automatic causal explanation analysis for combining arms training AAR. In Proceedings of the Industry/Interservice, Training, Simulation and Education Conference (I/ITSEC).

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. International Journal of Artificial Intelligence in Education, 10, 98 – 129.

Smith-Jentsch, K. A., Zeisig, R. L., Acton, B., McPherson, J. A. (1998). Team dimensional training: A strategy for guided team selfcorrection. In E. Salas and J. A. Cannon-Bowers (Eds). Making Decisions under Stress: Implications for Individual and Team Training, (pp. 271-297). Washington, D.C.: APA.

Stevens, S. M., Forsythe, J. C., Abbott, R. G. & Gieseler, C. J. (2009). Experimental assessment of accuracy of automated knowledge capture. Foundations of Augmented Cognition, Neuroergonomics, and Operational Neuroscience, 5638, 212-216.

Appendix I: NWDC Test Plan for EDRT AEMASE System

NWDC Modeling and Simulation Directorate
EDRT AEMASE Test 31Aug-2Sep
Integration Plan v1.1

Prepared For:

Modeling and Simulation
Navy Warfare Development Command

Prepared by:

John Zareno
john.zareno@nwdc.hpc.mil

28 July 2011

DOCUMENT CONTROL INFORMATION

Revision	Revision History	Date
V1.0	EDRT AEMASE ITP - Initial Draft	7/28/2011
V1.1	Updated test objectives	8/4/2011

TABLE OF CONTENTS

Integration Plan Scope

This Integration Plan provides the roadmap in evaluating the Automated Expert Modeling and Student Evaluation (AEMASE) module data capture interoperability and operational capabilities in support to the E-2C Deployable Readiness Trainer (EDRT) in a Navy Continuous Training Environment (NCTE) simulated Fleet Synthetic Training (FST) environment. Plan identifies the test resources, test environment, and test objectives to determine the level of NCTE compliance. Test Objectives will be accomplished by performing approved verification and validation methods as defined per test objective. Test objectives are designed to verify and identify capabilities and limitations of the tested application. Results of this test may identify further interoperability development requirements resulting in further integration testing. From here on, the EDRT AEMASE will be referred to as the data logger. Stimulation of simulation data will be performed using a Navy Warfare Development Command (NWDC) release version of Joint Semi Automated Forces (JSAF) approved for FST.

Referenced Documents

AEMASE Fleet Briefing Jul 2011

Test Objectives Summary

Basic Connectivity

Check that all training system devices are configured as per approved architecture. Verify proper indications for all network connections for all data streams to be tested (i.e. SIM, Tactical Voice, and Tactical Data Link). Verify data logger is configuration and that it properly joins federation.

Data Distribution Management (DDM)

Verify that data logger only subscribes to same DDM groups as training system and does not subscribe to DDM groups outside training systems interest.

Data Logging

Verify the data logger capabilities and limitations. Verify that data logger records all HLA tracks, simulated radio calls and radar tracks.

Schedule

Milestone	Date
Kickoff meeting	TBD
Test Readiness Review meeting	TBD
EDRT AEMASE Preliminary Inspection	31 Aug – 2 Sep

Resources and Configurations

Software Items

Table 5.1 Software

Software	Version	Purpose	Provider	Location

Table 5.1 Software

Software	Version	Purpose	Provider	Location
JSAF	JSAF v4.1.3.7 or subsequent release	NCTE Core Simulation	NWDC	NCTE Tier1
SAR	SAR RTI v2.1.2	Simulation Aware Router	NWDC	NCTE Tier 1 and Tier 3 nodes
Network Monitoring Tools	IP Traf, WireShark	Verify network network	NWDC	NCTE Tier1 and Tier 3 SAR Box
JBUS	v4.1.3.7 or subsequent release	Provide HLA/TADIL interface to GWM (Tier 1)	NWDC	NCTE Tier1
Training System	EDRT AEMASE *Build/Version*	Article Under Test	MFS	EDRT Norfolk Tier 3

Hardware Items

Table 0-1 is a list of hardware items that may be needed during integration:

Table 0-1 Hardware

Hardware	Version	Purpose	Provider	Location
Tier 1 Front End/Back Ends Machines	Std NCTE RHEL 5 image configured to support needed software (JSAF/JBUS)	Core Sim Hardware	NWDC	NCTE TIER 1
NCTE Tier 3 Node	Tier 3A	WAN/SAR equipment	NCTE OPS	VARIES
GWM	GM.020703.090508.1710.REL	Translates DIS into SIMPLE-J	NCTE MTT	NCTE TIER 1
Training System	EDRT, AEMASE	Article Under Test	CAE	VARIES
DIS-VOIP Gateway		DIS voice communications to VOIP protocol	NCTE	NCTE TIER 1

Federation Configuration Items

Table 0-2 Simulation Configuration Items

Simulation Technology	NCTE
HLA Federations	**NTF Federation** - NTF x.x FED and OMT, NTF.rid*, fedex *exercise specific name as defined in NTF.rid* The NTF federation will be configured to support latest NTF release approved for FST use.

Testing Architecture

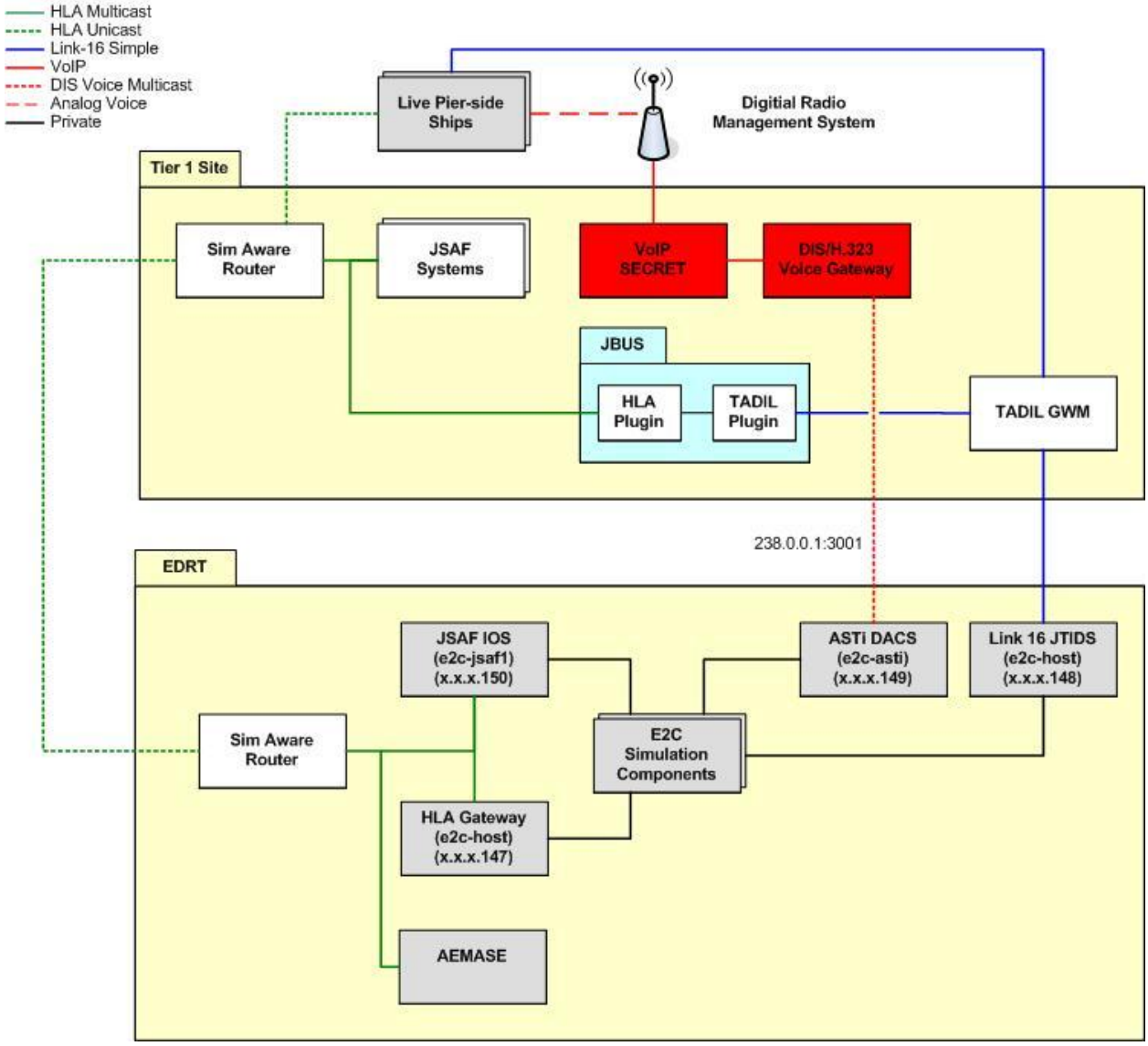

Table 0-3 EDRT SV-1

Test Procedures

Basic Connectivity

Test Objectives:

BC-1) Ping Connectivity – Can all training system devices be reached on NCTE Network?
BC-2) Federation Verification – Are all training system federates in the federation?
BC-3) Data Distribution Management – Are federates publishing/subscribing to correct routing space regions?
BC-4) Tactical Data Link Connectivity – Is the trainer in Link?
BC-5) Simulation Time Correlation – Does the training systems simulation time correlate with Zulu (UTC) time?

Preconditions:

- Tier 3 Node configured and operational
- Trainer Firewall has been opened to allow ping operations
- All federates are configured to Software Versions, Hardware, and Simulation Configuration Items in Table 0-1, Table 0-2 and **Error! Reference source not found.**
- Training system access switch port connecting to NCTE **enabled** and configured to **switch port access**.

Table 0-4 Basic Connectivity Objectives

Step	Test Method – Operator Action	Expected Response	Pass/Fail	Comments/Data Recorded
BC-1	Ping all trainer devices. Trainer ping Tier 3 Host default gateway.	Ping response	Pass	
BC-2	Verify trainer can join NTF federation (via NAVAIR DDM Proxy if applicable).	All federates are in the same federation	Pass	
BC-3	Verify trainer base multicast publication and subscriptions via netstat/wireshark.	Base Multicast groups should be within the range for fedex assigned as per RID	Pass	
BC-4	Verify training system is connected to the NCTE Gateway Manager.	NCTE Gateway Manager shows green indication for connectivity.	Pass	Test limited to basic connectivity and verification of PPLI and track exchange
BC-5	Verify training system's simulation time correlates with NCTE NTP server time in UTC/Zulu by marking several time points and object/interaction updates.	Information matches between training system and NCTE Time Server	Pass	

Data Distribution Management (DDM)

Test Objectives:
DDM-1) Verify data logger only subscribes to same DDM groups as training system
DDM-2) Verify bandwidth to/from node is within specification with all training system devices operational with no degradation in performance.

Pre-conditions:
- Tier 1 has robust scenario representative of actual FST event.

Table 0-5 Data Distribution Management

Step	Test Method – Operator Action	Expected Response	Pass/Fail	Comments/Data Recorded
DDM-1	Verify DDM subscriptions of Logger match those of the EDRT software by inspecting IP multicast groups joined	EDRT and logger join the same multicast groups	Pass	**Issue with JLCDT application when creating a private an adhoc fedex resolved by setting rid to use 1 MC group and interface to loopback**
DDM-2	Verify Logger does not cause increased bandwidth usage by measuring bytes sent/received from the wide area network which executing a scenario first without, then with the Logger.	Amount of network traffic is approximately equal with / without logger	Pass	**Issue stated above did cause significant increase in network traffic but was reolved with above work around.**

Data Logging

Test Objectives:

DL-1) Verify data logger records all HLA tracks
DL-2) Verify data logger records all simulated radio calls
DL-3) Verify data logger records all simulated radar tracks

Pre-conditions:

- Tier 1 has robust scenario representative of actual FST event.
- EDRT local network is configured to send radio traffic and mission computer data to logging host internal to EDRT.

Table 0-6 Data Distribution Management

Step	Test Method – Operator Action	Expected Response	Pass/Fail	Comments/Data Recorded
DL-1	Record scenario with the logger in two modes: 1) recording HLA entities 2) capturing only the statistics of the HLA entity updates (performance counters) to verify logger is not overwhelmed by the rate of updates received	The number of entity updates recorded should equal the performance counters from non-recording mode.	Pass	**4 hour recording captured**
DL-2	Run SQL query on recorded database to verify presence of radio calls.	All simulated radio calls are recorded. Intercoms are not recorded.		
DL-3	Examine Mission Computer logfile on capture node to verify track information is captured.	All track updates are captured.	[Sandia Note: this was subsequently verified]	

Testing Priority

Test objectives are listed in an order for which they should be executed. Individual test steps within each objective however may be completed out of order and does not necessarily prevent completion of remaining test steps.

Glossary/Acronym List

AEMASE – Automated Expert Modeling and Student Evaluation
DIS – Distributed Interactive Simulation
DDM - Data Distribution Management
EDRT – E-2C Deployable Readiness Trainer
FOM – Federation Object Model
FST – Fleet Synthetic Training
GWM – Gateway Manager
IOS – Instructor Operator System
JSAF – Joint Semi Automated Forces
NCTE – Navy Continuous Training Environment
PVD – Plain View Display
MCU – Multipoint Control Unit
NCTE – Navy Continuous Trainer Environment
NWDC – Navy Warfare Development Command
NTF – Navy Training Federation
NTP – Network Time Protocol
SAR – Simulation Aware Router
UTC – Coordinated Universal Time
VoIP – Voice over Internet Protocol